Basic Probability Theory: terminology

- An experiment has a set of potential outcomes, e.g., throw a dice, “look at” another sentence
- The sample space of an experiment is the set of all possible outcomes, e.g., \{1, 2, 3, 4, 5, 6\}
- In NLP our sample spaces tend to be very large
  - All words, bigrams, 5-grams
  - All sentences of length 20 (given a finite vocabulary)
  - All sentences
  - All parse trees over a given sentence

Corpus statistics

<table>
<thead>
<tr>
<th>Word</th>
<th>2008</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>government</td>
<td>7</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>dream</td>
<td>2</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>education</td>
<td>14</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>industry</td>
<td>5</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>innovation</td>
<td>0</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>win</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Afghanistan</td>
<td>2</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>Internet</td>
<td>0</td>
<td>D</td>
<td>E</td>
</tr>
</tbody>
</table>

Basic Probability Theory: terminology

- An event is a subset of the sample space
- Dice rolls
  - \( (2) \)
  - \( (3, 6) \)
  - even = \( \{2, 4, 6\} \)
  - odd = \( \{1, 3, 5\} \)
- NLP
  - a particular word/part of speech occurring in a sentence
  - a particular topic discussed (politics, sports)
  - sentence with a parasitic gap
  - pick your favorite phenomena…

Events

- We’re interested in probabilities of events
  - \( p(2) \)
  - \( p(\text{even}) \)
  - \( p(\text{odd}) \)
  - \( p(\text{parasitic gap}) \)
  - \( p(\text{word}) \)

Random variables

- A random variable is a mapping from the sample space to a number (think events)
- It represents all the possible values of something we want to measure in an experiment
- For example, random variable, \( X \), could be the number of heads for a coin

<table>
<thead>
<tr>
<th>space</th>
<th>HHH</th>
<th>HHT</th>
<th>HTH</th>
<th>HTT</th>
<th>THH</th>
<th>THT</th>
<th>TTH</th>
<th>TTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- Really for notational convenience, since the event space can sometimes be irregular

Random variables

- We can then talk about the probability of the different values of a random variable
- The definition of probabilities over all of the possible values of a random variable defines a probability distribution

<table>
<thead>
<tr>
<th>space</th>
<th>HHH</th>
<th>HHT</th>
<th>HTH</th>
<th>HTT</th>
<th>THH</th>
<th>THT</th>
<th>TTH</th>
<th>TTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P(X)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>( P(X=3) = 1/8 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>( P(X=2) = 3/8 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>( P(X=1) = 3/8 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>( P(X=0) = 1/8 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Probability distribution

- To be explicit
  - A probability distribution assigns probability values to all possible values of a random variable.
  - These values must be $\geq 0$ and $\leq 1$.
  - These values must sum to 1 for all possible values of the random variable.

<table>
<thead>
<tr>
<th>$X$</th>
<th>$P(X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>$P(X=3) = 1/2$</td>
</tr>
<tr>
<td>2</td>
<td>$P(X=2) = 1/2$</td>
</tr>
<tr>
<td>1</td>
<td>$P(X=1) = 1/2$</td>
</tr>
<tr>
<td>0</td>
<td>$P(X=0) = 1/2$</td>
</tr>
</tbody>
</table>

Unconditional/prior probability

- Simplest form of probability is $P(X)$.
- Prior probability: without any additional information, what is the probability?
  - What is the probability of a heads?
  - What is the probability of a sentence containing a pronoun?
  - What is the probability of a sentence containing the word “banana”?
  - What is the probability of a document discussing politics?
  - ...

Joint distribution

- We can also talk about probability distributions over multiple variables.
- $P(X, Y)$
  - probability of X and Y
  - a distribution over the cross product of possible values.

<table>
<thead>
<tr>
<th>NLPPass</th>
<th>P(NLPPass)</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>0.89</td>
</tr>
<tr>
<td>false</td>
<td>0.11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EngPass</th>
<th>P(EngPass)</th>
</tr>
</thead>
<tbody>
<tr>
<td>true</td>
<td>0.92</td>
</tr>
<tr>
<td>false</td>
<td>0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NLPPass AND EngPass</th>
<th>P(NLPPass, EngPass)</th>
</tr>
</thead>
<tbody>
<tr>
<td>true, true</td>
<td>0.88</td>
</tr>
<tr>
<td>true, false</td>
<td>0.01</td>
</tr>
<tr>
<td>false, true</td>
<td>0.04</td>
</tr>
<tr>
<td>false, false</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Joint distribution

- Still a probability distribution
  - all values between 0 and 1, inclusive
  - all values sum to 1
- All questions/probabilities of the two variables can be calculated from the joint distribution.

<table>
<thead>
<tr>
<th>NLPPass AND EngPass</th>
<th>P(NLPPass, EngPass)</th>
</tr>
</thead>
<tbody>
<tr>
<td>true, true</td>
<td>0.88</td>
</tr>
<tr>
<td>true, false</td>
<td>0.01</td>
</tr>
<tr>
<td>false, true</td>
<td>0.04</td>
</tr>
<tr>
<td>false, false</td>
<td>0.07</td>
</tr>
</tbody>
</table>

What is $P(\text{ENGPass})$?
Joint distribution

- Still a probability distribution
  - all values between 0 and 1, inclusive
  - all values sum to 1
- All questions/probabilities of the two variables can be calculated from the joint distribution

<table>
<thead>
<tr>
<th>NLPPass AND EngPass</th>
<th>P(NLPPass, EngPass)</th>
</tr>
</thead>
<tbody>
<tr>
<td>true, true</td>
<td>0.88</td>
</tr>
<tr>
<td>true, false</td>
<td>0.01</td>
</tr>
<tr>
<td>false, true</td>
<td>0.04</td>
</tr>
<tr>
<td>false, false</td>
<td>0.07</td>
</tr>
</tbody>
</table>

0.92

How did you figure that out?

Joint distribution

\[ P(x) = \sum_{y \in Y} p(x, y) \]

Conditional probability

- As we learn more information, we can update our probability distribution
- \( P(X \mid Y) \) models this (read “probability of \( X \) given \( Y \)”):
  - What is the probability of a heads given that both sides of the coin are heads?
  - What is the probability the document is about politics, given that it contains the word “Clinton”?
  - What is the probability of the word “banana” given that the sentence also contains the word “split”?
- Notice that it is still a distribution over the values of \( X \)

Conditional probability

\[ p(X \mid Y) = ? \]

In terms of prior and joint distributions, what is the conditional probability distribution?
Conditional probability

\[ p(X | Y) = \frac{P(X, Y)}{P(Y)} \]

Given that \( y \) has happened, in what proportion of those events does \( x \) also happen?

What is: \( p(\text{NLPPass} = \text{true} | \text{EngPass} = \text{false}) \)?

<table>
<thead>
<tr>
<th>NLPPass AND EngPass</th>
<th>P(NLPPass, EngPass)</th>
</tr>
</thead>
<tbody>
<tr>
<td>true, true</td>
<td>.88</td>
</tr>
<tr>
<td>true, false</td>
<td>.01</td>
</tr>
<tr>
<td>false, true</td>
<td>.04</td>
</tr>
<tr>
<td>false, false</td>
<td>.07</td>
</tr>
</tbody>
</table>

\[ P(\text{EngPass} = \text{false}) = 0.01 + 0.07 = 0.08 \]

\[ P(\text{true, false}) = 0.01 \]

Notice this is very different than \( p(\text{NLPPass} = \text{true}) = 0.89 \)

A note about notation

- When talking about a particular assignment, you should technically write \( p(X=x) \), etc.
- However, when it’s clear, we’ll often shorten it.
- Also, we may also say \( P(X) \) or \( p(x) \) to generically mean any particular value, i.e. \( P(X=x) \)

\[ P(\text{true, false}) = 0.01 \]

\[ P(\text{EngPass} = \text{false}) = 0.01 + 0.07 = 0.08 \]

\[ = 0.125 \]
Properties of probabilities

- $\mathbb{P}(A \text{ or } B) = \, ?$

Properties of probabilities

- $\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A, B)$

Properties of probabilities

- $\mathbb{P}(\neg E) = 1 - \mathbb{P}(E)$
- More generally:
  - Given events $E = e_1, e_2, \ldots, e_n$
    \[ p(e_i) = 1 - \sum_{j=1; j\neq i}^n p(e_j) \]
  - $\mathbb{P}(E_1, E_2) \leq \mathbb{P}(E_1)$

Properties of probabilities

- Chain rule (aka product rule)

  \[ p(X \mid Y) = \frac{p(X, Y)}{P(Y)} \quad \rightarrow\quad p(X, Y) = P(X \mid Y)P(Y) \]

  We can view calculating the probability of $X$ AND $Y$
  occurring as two steps:
  1. $Y$ occurs with some probability $P(Y)$
  2. Then, $X$ occurs, given that $Y$ has occurred

or you can just trust the math... 😊
Chain rule

\[ p(X,Y,Z) = P(X,Y,Z)p(Y,Z) \]
\[ p(X,Y,Z) = P(X,Y|Z)p(Z) \]
\[ p(X,Y,Z) = P(Y|Z)p(Y)p(X) \]

\[ p(X_1,X_2,...,X_n) = ? \]

Applications of the chain rule

- We saw that we could calculate the individual prior probabilities using the joint distribution
  \[ p(x) = \sum_{y \in Y} p(x,y) \]
- What if we don't have the joint distribution, but do have conditional probability information:
  - \( P(Y) \)
  - \( P(X|Y) \)
  \[ p(x) = \sum_{y \in Y} p(y)p(x|y) \]

Bayes’ rule (theorem)

\[ p(X|Y) = \frac{P(X,Y)}{P(Y)} \]
\[ p(X,Y) = P(X|Y)p(Y) \]

\[ p(Y|X) = \frac{P(X,Y)}{P(X)} \]
\[ p(X,Y) = P(Y|X)p(X) \]

\[ p(x|y) = \frac{P(Y|X)p(X)}{P(Y)} \]

Bayes’ rule

- Allows us to talk about \( P(Y|X) \) rather than \( P(X|Y) \)
- Sometimes this can be more intuitive
- Why?

\[ p(X|Y) = \frac{P(Y|X)p(X)}{P(Y)} \]
Bayes rule

- $p(\text{disease} \mid \text{symptoms})$
  - For everyone who had those symptoms, how many had the disease?
- $p(\text{symptoms} \mid \text{disease})$
  - For everyone that had the disease, how many had this symptom?

- $p(\text{linguistic phenomena} \mid \text{features})$
  - For all examples that had those features, how many had that phenomena?
- $p(\text{features} \mid \text{linguistic phenomena})$
  - For all the examples with that phenomena, how many had this feature

- $p(\text{cause} \mid \text{effect})$ vs. $p(\text{effect} \mid \text{cause})$

Gaps

I just won’t put these away.

These, I just won’t put away.

I just won’t put ___ away.

What did you put ___ away?

The socks that I put ___ away.

Whose socks did you fold ___ and put ___ away?

Whose socks did you fold ___?

Whose socks did you put ___ away?
Parasitic gaps

These I'll put _____ away without folding _____.

These I'll put _____ away.

These without folding _____.

Parasitic gaps

1. Cannot exist by themselves (parasitic)
   These I'll put my pants away without folding _____.

2. They’re optional
   These I'll put ____ away without folding them.

Parasitic gaps


Frequency of parasitic gaps

- Parasitic gaps occur on average in 1/100,000 sentences
- Problem:
  - Joe Linguist has developed a complicated set of regular expressions to try and identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn’t, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it is a parasitic gap, what is the probability it actually is?
Joe Linguist has developed a complicated set of regular expressions to try and identify parasitic gaps. If a sentence has a parasitic gap, it correctly identifies it 95% of the time. If it doesn’t, it will incorrectly say it does with probability 0.005. Suppose we run it on a sentence and the algorithm says it is a parasitic gap, what is the probability it actually is?

What question do we want to ask?

\[
p(g | t) = \frac{p(t | g)p(g)}{p(t)} \]

\[
= \frac{p(t | g)p(g)}{\sum_{g \in G} p(g)p(t | g)} \]

\[
= \frac{p(t | g)p(g)}{p(g)p(t | g) + p(\overline{g})p(t | \overline{g})} \]

\[
= \frac{0.95 \times 0.00001}{0.00001 \times 0.95 + 0.99999 \times 0.005} \approx 0.002
\]
Obtaining probabilities

- We’ve talked a lot about probabilities, but not where they come from.
  - What is the probability of “the” occurring in a sentence?
  - What is the probability of “Middlebury” occurring in a sentence?
  - What is the probability of “I think today is a good day to be me” as a sentence?

Estimating probabilities

- What is the probability of “the” occurring in a sentence?
  
  We don’t know!

  We can estimate that based on data, though:

  \[
  \frac{\text{number of sentences with “the”}}{\text{total number of sentences}}
  \]

Maximum likelihood estimation

- Intuitive
- Sets the probabilities so as to maximize the probability of the training data

- Problems?
  - Amount of data
  - Particularly problematic for rare events
  - Is our training data representative
Say the actual probability is 1/100,000

We don't know this, though, so we're estimating it from a small data set of 10K sentences

What is the probability that, by chance, we have a parasitic gap sentence in our sample?

\[ P(\text{not parasitic}) = 0.99999 \]

\[ P(\text{not parasitic})^{10000} \approx 0.905 \] is the probability of us NOT finding one

So, probability of us finding one is \( \sim 10\% \), in which case we would incorrectly assume that the probability is 1/10,000 (10 times too large)

Remember Zipf's law from last time... NLP is all about rare events!